

**OPEN ACCESS**

EDITED BY  
Alexandre Hudon,  
University of Montreal, Canada

REVIEWED BY  
Laurent Elkrief,  
Montreal University, Canada  
Mohammad Hossein Salemi,  
University of Tehran, Iran

\*CORRESPONDENCE  
Eik Niederlohmann  
✉ kontakt@praxis-niederlohmann.de

RECEIVED 18 November 2025  
REVISED 09 February 2026  
ACCEPTED 11 February 2026  
PUBLISHED 13 March 2026

CITATION  
Niederlohmann E (2026) Perceive–  
Assess–Dose–Safeguard: a safety-gated  
state–action grammar for  
psychotherapy micro-decisions in  
computational psychiatry.  
*Front. Psychiatry* 17:1749364.  
doi: 10.3389/fpsy.2026.1749364

COPYRIGHT  
© 2026 Niederlohmann. This is an open-  
access article distributed under the terms  
of the [Creative Commons Attribution  
License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or  
reproduction in other forums is  
permitted, provided the original  
author(s) and the copyright owner(s) are  
credited and that the original publication  
in this journal is cited, in accordance  
with accepted academic practice. No  
use, distribution or reproduction is  
permitted which does not comply with  
these terms.

# Perceive–Assess–Dose– Safeguard: a safety-gated state– action grammar for psychotherapy micro-decisions in computational psychiatry

Eik Niederlohmann\*

Department of Psychosomatic Medicine and Psychotherapy, Kliniken Erlabrunn,  
Breitenbrunn, Germany

Psychotherapy unfolds as a sequence of rapid micro-decisions under uncertainty. Within seconds, clinicians integrate verbal, paraverbal, embodied, and relational cues, estimate the patient's momentary capacity for affective work, choose an intervention dose, and apply stop rules to prevent overwhelm and rupture. Computational psychiatry offers principled frameworks for sequential decision-making, but progress in computational psychotherapy remains constrained by the lack of clinically grounded, machine-readable grammars that capture therapist micro-decisions in context. I introduce the Perceive-Assess-Dose-Safeguard (PAD-S) decision matrix as a safety-gated state–action grammar for psychotherapy micro-decisions. PAD-S formalizes four “front-of-system” signals—defensive/avoidant organization (DEF), anxiety/arousal and tolerance (ANX), patient progression toward direct experience and action (PRO), and self-attack/shame processes (SUP)—together with three safety thresholds (A–C) that gate intervention dose. Each decision point can be logged as an “episode line” (trigger, state, threshold, action, and expected functional impact), enabling transcript annotation and structured datasets. PAD-S is grounded in experiential dynamic psychotherapy (EDT/ISTDP) yet expressed as an orientation-translatable representation layer: DEF can be read as avoidance/safety behavior, ANX as arousal/tolerance, PRO as approach and value-consistent action, and SUP as self-criticism/shame. I show how PAD-S trajectories can interface with hybrid neural–cognitive models such as SPICE to discover sparse, interpretable equations of process change, and I outline testable hypotheses and feasible pilot studies (reliability, outcome linkage, and modeling) to evaluate the framework. computational psychiatry; computational psychotherapy; psychotherapy process coding; interpretable AI; human-in-the-loop; active inference; SPICE; Mini-ICF-APP

**KEYWORDS**

active inference, computational psychiatry, computational psychotherapy, human-in-the-loop, interpretable AI, Mini-ICF-APP, psychotherapy process coding, SPICE

## 1 Introduction

Computational psychiatry has advanced mechanistic accounts of learning and decision-making, and related work on computational psychological therapy has argued that formal models can inform psychotherapy training and personalization (1–3). Yet psychotherapy is not a laboratory task: it unfolds in time, in context, and within a relationship, often under strong affect and uncertainty—precisely the conditions that computational psychiatry has identified as essential to model explicitly (4).

A central bottleneck for computational psychotherapy is representational. Many computational approaches can learn from sequential data, but psychotherapy transcripts rarely come with clinically meaningful, machine-readable labels of the therapist's moment-to-moment decisions. A recent scoping review of computational methods in psychotherapy highlights substantial heterogeneity in how process is operationalized, coded, and linked to outcomes (5). In parallel, large language models are increasingly used for psychotherapy-related tasks, which raises both opportunities and governance challenges; however, their utility still depends on clear target representations and high-quality supervision signals (6).

Hybrid neural-cognitive approaches can mitigate the interpretability gap by combining flexible sequence models with equation discovery. For example, SPICE automates the discovery of sparse and interpretable cognitive equations from sequential behavior (7). Such approaches have also been discussed in the broader context of automating parts of the scientific workflow, including hypothesis generation and model selection (8). To bring these tools to psychotherapy, we need a compact state-action grammar that is clinically grounded, safety-aware, and feasible to annotate.

The Perceive-Assess-Dose-Safeguard (PAD-S) decision matrix is proposed as such a grammar. PAD-S is theoretically grounded in experiential dynamic psychotherapy traditions (EDT/ISTDP), which emphasize real-time monitoring of defensive organization, anxiety/arousal tolerance, patient progression, and self-attack/shame processes to titrate the dose of affective work and protect the therapeutic relationship (9–12). PAD-S grew out of an earlier transdiagnostic “Conflict Square Algorithm” that aimed to support functional formulation and documentation (13), but it is refined here as an explicit state-action representation for transcript annotation and computational modeling.

Importantly, PAD-S is not presented as “school-neutral” in the sense of being theory-free. Instead, it is an orientation-translatable intermediate representation: its state variables can be mapped to constructs familiar in multiple traditions (e.g., DEF as avoidance or safety behavior; ANX as arousal and tolerance; PRO as approach, emotional access, and value-consistent action; SUP as self-criticism and shame). This translation stance is intended to make PAD-S usable across modalities without requiring endorsement of psychodynamic metatheory, while still preserving the clinically pragmatic heuristics that motivated the model.

PAD-S is also designed to complement—not replace—existing psychotherapy process measurement traditions. Technique-specific

coding systems (e.g., motivational interviewing skill codes, CBT adherence/competence scales), Q-set approaches (e.g., psychotherapy process Q-set), and relational theme methods (e.g., core conflictual relationship themes) capture important dimensions of process. PAD-S adds a safety-gated decision layer that explicitly represents the therapist's momentary assessment of tolerance and the consequent dose/stop-rule choices, which can be integrated with richer descriptive systems when needed (5).

Figure 1 provides an overview of PAD-S as (i) a clinical micro-decision workflow and (ii) a data pipeline that converts transcript episodes into structured state-action trajectories amenable to downstream modeling. Box 1 provides a minimal formalization framing PAD-S as a safety-constrained policy over discrete states.

## 2 The PAD-S decision matrix

PAD-S is a compact decision matrix that represents psychotherapy micro-decisions as a sequence of safety-gated state-action steps. At each decision point, the therapist (i) perceives clinically relevant cues, (ii) assesses the current “front-of-system” state and tolerance threshold, (iii) selects an intervention dose and type, and (iv) safeguards the process with explicit stop rules when risk markers indicate overload or shame-related rupture risk (Figure 1). Figure 2 illustrates a minimal PAD-S episode line representation for transcript-based modeling.

### 2.1 Perceive: cues that matter for micro-decisions

Perception in PAD-S refers to identifying momentary cues that are actionable for process management. These cues can be verbal (content, contradictions, commitment language), paraverbal (tempo, prosody), embodied (tension patterns, breath, gaze), and relational (alliance signals, ruptures, responsiveness). PAD-S does not assume that any single channel is privileged; rather, cues serve as proxies for the current state of affect regulation and action readiness that inform dosing and safeguarding.

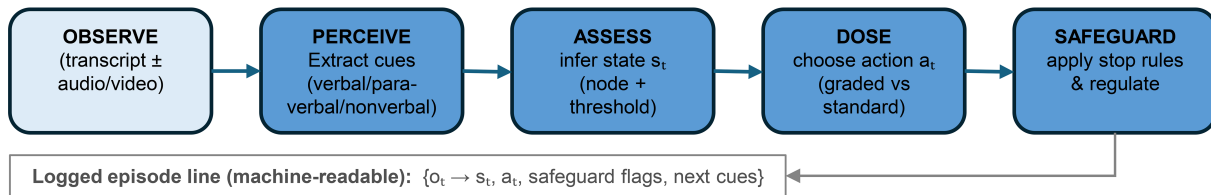
### 2.2 Assess: four front-of-system nodes and three tolerance thresholds

The core state representation in PAD-S combines a categorical node (DEF, ANX, PRO, SUP) with a tolerance threshold (A, B, C). Nodes summarize what dominates the process in the moment; thresholds gate how much affective “pressure” or exposure can be safely applied without pushing the system into overload.

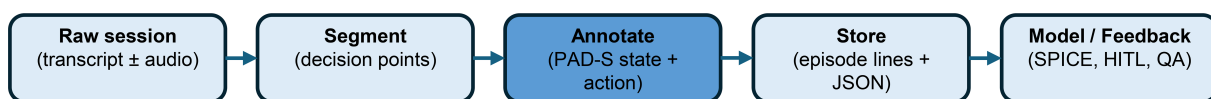
Nodes (front-of-system signals): DEF captures defensive or avoidant organization (e.g., detours, intellectualization, reassurance seeking, disengagement). ANX captures anxiety/arousal and tolerance (e.g., somatic tension, dysregulated arousal, cognitive-perceptual disruption). PRO captures progression: direct access to feeling, clarity of wish/goal, and readiness for value-consistent action. SUP captures self-attack/shame processes (e.g., self-criticism, moralizing collapse, joy-to-attack switches). While these labels are grounded in EDT/ISTDP,

### PAD-S as a safety-gated state-action interface (human-final → model-ready)

#### A In-session micro-decision grammar (OBSERVE → PERCEIVE → ASSESS → DOSE → SAFEGUARD)



#### B Episode-line pipeline for computational teams (LLMs / SPICE / QA; clinician adjudicates)



Goal: a compact, interpretable state-action trajectory that supports human-in-the-loop supervision and safe AI development.

FIGURE 1

PAD-S as a safety-gated state-action interface (human-final → model-ready). (A) In-session micro-decision grammar. PAD-S is represented as a safety-gated loop: OBSERVE (transcript ± audio/video) → PERCEIVE clinically actionable cues (verbal/paraverbal/nonverbal) → ASSESS the current interaction state  $s_t$  (front-of-system node + tolerance threshold) → DOSE the next-step intervention/action  $a_t$  (graded vs. standard) → SAFEGUARD by applying explicit stop rules and regulation when overload/shame-collapse markers emerge. Each decision point is logged as a machine-readable episode line (e.g.,  $o_t \rightarrow s_t, a_t$ , safeguard flags, next cues). (B) Episode-line pipeline for computational teams. Raw sessions are segmented into clinically meaningful decision points, annotated with PAD-S state + action labels, stored as episode lines (e.g., JSON), and converted into state-action trajectories for downstream modeling (e.g., interpretable hybrid approaches) and human-in-the-loop feedback/quality assurance, while keeping the clinician in human-final control.

they can be translated into common constructs used across modalities (avoidance, arousal tolerance, approach/action, self-criticism).

Thresholds (A-C): Threshold A indicates a regulated window in which standard interventions are typically tolerated. Threshold B indicates heightened arousal or fragility requiring graded dosing, shorter exposures, and more frequent regulation checks. Threshold C indicates overload markers (e.g., cognitive-perceptual disruption, panic-level dysregulation, or shame collapse), requiring immediate safeguarding (stop deepening, regulate, protect positives, repair alliance) before any further exposure or pressure is attempted.

### 2.3 Dose and safeguard: action selection under safety constraints

Dose refers to the intensity and format of the next-step intervention (e.g., standard vs. graded), as well as the choice of node-appropriate actions (e.g., brief defense-blocking vs. supportive reflection; affect focus vs. grounding; protect positives vs. interpretation). Safeguard refers to explicit stop rules that prioritize safety and alliance: if C-level markers appear, the policy requires shifting from deepening to regulation and repair.

#### BOX 1 Minimal formalization of PAD-S as a safety-gated state-action policy.

Unit of analysis: decision point/episode  $t$  within a session.

Observations:  $o_t =$  cues perceived by the therapist (verbal, paraverbal, embodied, relational).

State:  $s_t = (\text{node}_t, \text{thr}_t, \text{ctx}_t)$ , where  $\text{node}_t \in \{\text{DEF}, \text{ANX}, \text{PRO}, \text{SUP}\}$  and  $\text{thr}_t \in \{A, B, C\}$ .

Actions:  $a_t \in A$ , where  $A$  includes node-appropriate interventions (e.g., clarify/invite, block a defense, focus affect, regulate, protect positives, repair alliance).

Dose format:  $f_t \in \{\text{standard or graded}\}$ .

Safety constraint (stop rule): if  $\text{thr}_t = C$ , restrict actions to safeguarding and regulation until  $\text{thr}_t$  returns to A or B.

Episode line (log record):  $l_t = (\text{trigger}_t, s_t, a_t, f_t, \text{expected\_function}_t)$ .

Structured dataset:  $D = \{l_t\}$  for  $t = 1..T$  across sessions/cases.

Pseudo-code (high-level):

- 1)  $o_t \leftarrow \text{observe}()$
- 2)  $\text{node}_t \leftarrow \text{classify\_front\_of\_system}(o_t)$
- 3)  $\text{thr}_t \leftarrow \text{estimate\_tolerance}(o_t)$
- 4) if  $\text{thr}_t = C$ :  $a_t \leftarrow \text{safeguard\_and\_regulate}(\text{node}_t)$
- else:  $a_t \leftarrow \text{choose\_next\_step}(\text{node}_t, \text{thr}_t)$  (standard or graded)
- 5)  $l_t \leftarrow \text{log}(\text{trigger}_t, \text{node}_t, \text{thr}_t, a_t, f_t, \text{expected\_function}_t)$
- 6) proceed to next decision point

## Minimal PAD-S episode line (example) for transcript-based modeling

Input: transcript decision point (anonymized)

T: "When you say it's fine, what happens inside right now?"  
 P: "Nothing... I should just handle it." (pause; quieter voice; gaze down)

Output: PAD-S episode line (state + action)

Node	Thr.	Dose (a <sub>i</sub> )	Safe ?	Notes
DEF	B	graded focus + regulation	yes	avoidance as feedback; return to felt sense

Example JSON:  

```
{"node": "DEF", "thr": "B", "dose": "graded", "safe": true}
```

Sequence of episode lines → state–action trajectory → interpretable sequential models (e.g., SPICE) + human-in-the-loop feedback

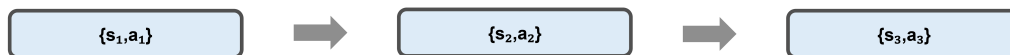


FIGURE 2

Minimal PAD-S episode line (example) for transcript-based modeling. A single transcript decision point (T = therapist; P = patient; anonymized illustrative example) is mapped to a compact PAD-S episode line capturing the assessed Node (front-of-system signal), Thr. (tolerance threshold), selected intervention Dose/Action (a<sub>i</sub>: graded vs. standard), a Safeguard indicator (whether stop-rule/regulation was applied), and brief notes. The episode line can be stored as a structured record (e.g., JSON) and concatenated across decision points to form state–action trajectories ((s<sub>1</sub>, a<sub>1</sub>), (s<sub>2</sub>, a<sub>2</sub>), ...), enabling interpretable sequential modeling and human-in-the-loop supervision/feedback.

This safety-gated design is intended to make the model usable for computational purposes without implying automation of clinical decisions. PAD-S is a descriptive and prescriptive grammar for human decision-making: it makes the clinician’s implicit gating logic explicit and recordable.

### 2.4 The episode line: a minimal unit for annotation and functional linkage

To support transcript annotation and downstream modeling, PAD-S uses the “episode line” as a minimal recording unit. An episode line captures (i) a trigger/cue, (ii) the assessed node(s), (iii) the threshold (A-C), (iv) the selected action and dose format, and (v) the expected functional impact (e.g., on planning, endurance, social interaction, self-care). This functional orientation can be linked to standardized domains such as the Mini-ICF-APP, a widely used instrument for assessing activity and participation limitations in mental disorders (14). The episode line may also include a planned re-check interval; this refers to when the specified functional target (e.g., a Mini-ICF-APP domain) should be reassessed in routine care (for example after ~4–6 weeks of weekly sessions) rather than to the timing of session-by-session coding.

Supplementary Materials S1-S3 provide an implementable codebook, a decision matrix, annotated examples, a JSON schema for episode lines, and a cross-walk to functional domains. Box 1 summarizes a minimal computational formalization of PAD-S that is sufficient to situate the framework within control-theoretic and hybrid-modeling perspectives.

State and observations. At each decision point *t*, PAD-S uses observable cues *o<sub>t</sub>* (verbal, paraverbal, embodied, relational) to approximate a latent interaction state *s<sub>t</sub>* = (node<sub>*t*</sub>, thr<sub>*t*</sub>, ctx<sub>*t*</sub>), where node<sub>*t*</sub> ∈ {DEF, ANX, PRO, SUP} and thr<sub>*t*</sub> ∈ {A, B, C}.

Actions. The clinician selects a next-step action *a<sub>t</sub>* from a small, interpretable set *A* (e.g., clarify\_defense, regulate\_ground, graded\_exposure, standard\_exposure, protect\_positives, alliance\_repair).

Policy. PAD-S defines a safety-gated policy π<sub>PAD-S</sub>(*a<sub>t</sub>*|*s<sub>t</sub>*) implemented as a node × threshold lookup rule (Supplementary Material S1). A hard safety constraint applies: if thr<sub>*t*</sub> = C (cognitive–perceptual disruption or shame collapse), deepening actions are disallowed and the action must prioritize regulation/repair (and protection of positives when relevant).

Episode line. Each decision point is recorded as *e<sub>t</sub>* = (trigger<sub>*t*</sub>, *s<sub>t</sub>*, *a<sub>t</sub>*, expected\_function<sub>*t*</sub>, recheck<sub>*t*</sub>). The resulting trajectory {*e<sub>t</sub>*} provides discrete, human-interpretable labels for downstream modeling (e.g., hybrid neural–cognitive approaches; Section 4) and for training/supervision feedback.

Minimal pseudocode:  
 observe cues *o<sub>t</sub>*.  
 infer node<sub>*t*</sub>, thr<sub>*t*</sub>.  
 if thr<sub>*t*</sub> == C: *a<sub>t</sub>* ← regulate/repair (+ protect\_positives if needed).  
 else: *a<sub>t</sub>* ← lookup(node<sub>*t*</sub>, thr<sub>*t*</sub>).  
 log episode line *e<sub>t</sub>*.

### 3 PAD-S as a control problem: resource-rational and active-inference perspectives

PAD-S can be read as a high-level control policy for a coupled human system (patient–therapist dyad) operating under uncertainty and safety constraints. At each decision point, the clinician performs approximate state estimation (node and threshold) and selects an

action intended to move the process toward progression (PRO) while avoiding catastrophic states (e.g., overload or shame collapse).

From a resource-rational perspective, therapists face computational constraints: they must allocate attention, working memory, and inference to the most informative cues and actions in real time. Rational metareasoning models and resource-rational analysis provide a principled language for why therapists may prefer simple heuristics and why dosing and safeguarding functions are essential under limited cognitive control resources (15, 16). PAD-S makes these heuristics explicit and thus testable.

From an active-inference perspective, PAD-S corresponds to (i) selecting observations that reduce uncertainty (Perceive), (ii) inferring a latent state of the interaction (Assess), (iii) choosing a policy (Dose) that trades off epistemic and pragmatic value, and (iv) enforcing safety priors that prevent high-cost prediction errors (Safeguard) (17). Recent work extending active inference toward social actors and context-sensitive regulation provides a natural conceptual bridge for modeling dyadic psychotherapy as an embedded process rather than an isolated decision-maker (18).

## 4 Mapping PAD-S to hybrid neural–cognitive models such as SPICE

PAD-S produces discrete state–action trajectories that are well suited to hybrid modeling. Each episode line can be interpreted as a time step  $t$  with state  $s_t = (\text{node}_t, \text{thr}_t, \text{context}_t)$  and action  $a_t$  (Box 1). Sequences of episode lines across a session form trajectories that encode how therapist actions interact with patient state transitions (e.g., DEF  $\rightarrow$  ANX  $\rightarrow$  PRO, or PRO  $\rightarrow$  SUP).

SPICE combines sequence modeling with sparse equation discovery to recover interpretable structural relationships from behavioral trajectories (7). In psychotherapy, PAD-S trajectories can serve as the observable scaffold: SPICE (or related methods) can be applied to discover compact equations that predict transitions between nodes and thresholds as a function of prior state, therapist action, and context. This could yield testable mechanistic hypotheses, such as whether particular action classes increase the probability of PRO under B-level thresholds, or whether “protect positives” actions reduce the probability of SUP following PRO in vulnerable patients.

A key advantage of coupling PAD-S with equation discovery is interpretability. Rather than treating the therapy session as a black-box sequence, PAD-S constrains the state space to clinically meaningful variables and safety gating, which can reduce the risk of overfitting and improve cross-setting transportability. Figure 1B summarizes this interface from transcript episodes to structured datasets and hybrid modeling outputs.

## 5 Research agenda: testable hypotheses and feasible pilot designs

PAD-S is intended as a hypothesis-generating and hypothesis-testing bridge between clinical process knowledge and

computational models. Below, I outline example hypotheses and practical pilot designs that can be executed with modest resources and that directly address the feasibility, reliability, and scientific value of PAD-S annotation.

### 5.1 Example testable hypotheses

1. H1 (policy safety adherence): Across therapists, C-level threshold markers will be followed predominantly by safeguarding actions (regulation/repair). Higher safety adherence (fewer deepening actions at C) will be associated with fewer alliance ruptures and lower dropout.
2. H2 (dose–tolerance matching): Therapists will preferentially use graded dosing under B-level thresholds and standard dosing under A-level thresholds. Greater dose–tolerance matching will predict functional gains over treatment.
3. H3 (protect-positives mechanism): In patients with prominent self-criticism/shame, PRO episodes will be followed by SUP more often unless the therapist implements “protect positives” actions. Protect-positives actions will reduce the probability of PRO  $\rightarrow$  SUP transitions.
4. H4 (model discovery and responder signatures): Applying equation discovery (e.g., SPICE) to PAD-S trajectories will yield sparse transition equations. Responders vs. nonresponders will show distinct structural patterns (e.g., stronger DEF attractors, higher SUP reactivity, or reduced sensitivity to therapist actions) that can be tested prospectively.

### 5.2 Feasible pilot study designs

The following pilots are structured to (i) establish annotation reliability, (ii) link PAD-S process metrics to functional outcomes, and (iii) test the value of hybrid modeling on PAD-S-labeled data.

Pilot 1: Annotation feasibility and interrater reliability.

- Sample: 10–20 de-identified session transcripts (diverse modalities and case mixes).
- Raters: 3–5 trained raters; iterative calibration with a shared rater pack.
- Unit: episode line segmentation at clinically meaningful decision points (rather than continuous second-by-second coding).
- Outcomes: interrater reliability for node (DEF/ANX/PRO/SUP) and threshold (A/B/C). Target:  $\kappa \geq 0.70$  for node and threshold; ICC for any continuous ratings if used.
- Deliverable: refined codebook and adjudication rules; estimate of annotation time per session.

Pilot 2: Functional linkage study.

- Sample: ~30–60 patients with baseline and follow-up functional assessment (e.g., Mini-ICF-APP).
- Measures: derive PAD-S process metrics per case (e.g., proportion of PRO time, frequency of C markers, PRO  $\rightarrow$  SUP transitions, dose–tolerance matching index).

- Hypothesis test: regress functional change on PAD-S metrics while controlling for baseline severity; focus on effect sizes and feasibility rather than definitive inference.

Pilot 3: Hybrid modeling/equation discovery on PAD-S trajectories.

- Data: PAD-S episode-line trajectories from Pilot 2 (or larger corpus if available).
- Modeling: fit a sequence model to capture dynamics; apply sparse regression/equation discovery (e.g., SPICE) to recover interpretable transition equations.
- Evaluation: predictive generalization (held-out sessions/cases), stability of discovered equations across bootstraps, and clinical interpretability (expert review).
- Deliverable: candidate mechanistic equations and responder/nonresponder signatures for prospective testing.

### 5.3 Feasibility demonstration: proof-of-concept episode-line annotation

Feasibility benefits from a concrete example. Below is a short proof-of-concept that illustrates how PAD-S translates a brief therapist–patient exchange into one episode line and a minimal JSON record. The excerpt is a fictional composite created for training and illustration; it does not reproduce copyrighted or identifiable clinical material.

Therapist: “When your manager criticized you, what happened inside—right now as you remember it?”

Patient: “My stomach clenched and I went blank. I told him it’s fine.”

Episode line (human-readable): Trigger=manager criticism → Node=ANX (with DEF detour) → Threshold=B (narrowing; smooth-muscle activation) → Action=downshift + grounding + graded micro-focus → Functional target=endurance/persistence & planning/structuring → Re-check=4–6 weeks.

JSON example: {"trigger": "manager criticism", "node": "ANX", "secondary\_node": "DEF", "threshold": "B", "action": ["grounding", "graded\_focus"], "function\_target": ["endurance", "planning"], "recheck\_weeks": 6}.

## 6 Discussion

PAD-S is proposed as a compact, safety-gated state–action grammar for psychotherapy micro-decisions. It translates pragmatic process heuristics from experiential dynamic psychotherapy into a form that can be (i) annotated in transcripts, (ii) logged in clinical documentation as episode lines, and (iii) used as an interface to hybrid modeling approaches such as equation discovery. The framework is intentionally minimal: it aims to capture a decision layer (node, threshold, action) that can be

combined with richer descriptive coding systems when research questions require finer granularity.

Several limitations are important. First, PAD-S is grounded in a specific clinical tradition (EDT/ISTDP) and therefore inherits its assumptions about what cues and interventions matter. The translation stance described here is a pragmatic proposal, not an empirical claim of equivalence between modalities. Second, PAD-S is not a substitute for clinical judgment; it formalizes heuristics and safety gating but cannot capture the full nuance of case formulation, ethics, or relational context.

Feasibility and reliability are central. Episode-line coding reduces cognitive load compared with continuous micro-coding, but it still requires training, calibration, and transparent adjudication rules. Supplementary rater materials can support a calibration workflow in which raters first converge on segmentation rules and then on node/threshold assignments. In practice, reliability will likely vary by node (e.g., PRO and SUP may require more context) and by setting. Reporting annotation time,  $\kappa$ /ICC, and failure modes should be treated as primary outcomes in early studies.

Finally, governance and safety must remain explicit. PAD-S should be used to support human learning (training, supervision, hypothesis testing), not to automate treatment decisions. If machine learning is used for pre-annotation or decision support, therapists must retain human-final control, and privacy-preserving workflows are required (19).

## Data availability statement

The original contributions presented in the study are included in the article/[Supplementary Material](#). Further inquiries can be directed to the corresponding author/s.

## Author contributions

EN: Conceptualization, Methodology, Writing – original draft, Writing – review & editing.

## Funding

The author(s) declared that financial support was not received for this work and/or its publication.

## Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Generative AI statement

The author(s) declared that generative AI was used in the creation of this manuscript. During manuscript preparation, the author used ChatGPT (OpenAI; ChatGPT Plus; model: GPT-5.2; accessed February 2026) to support language editing (clarity, grammar, and readability) and literature exploration. No identifiable or sensitive patient data were entered into the tool. All AI-assisted outputs were critically reviewed, verified, and edited by the author, who remains fully responsible for the originality, accuracy, and integrity of the manuscript, including all citations and references.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpsy.2026.1749364/full#supplementary-material>

## References

- Nair A, Rutledge RB, Mason L. Under the hood: using computational psychiatry to make psychological therapies more mechanism-focused. *Front Psychiatry*. (2020) 11:140. doi: 10.3389/fpsy.2020.00140
- Moutoussis M, Shahar N, Hauser TU, Dolan RJ. Computation in psychotherapy, or how computational psychiatry can aid learning-based psychological therapies. *Comput Psychiatr*. (2018) 2:50–73. doi: 10.1162/CPSY\_a\_00014
- Berwian IM, Hitchcock PF, Pisupati S, Schoen G, Niv Y. Using computational models of learning to advance cognitive behavioral therapy. *Commun Psychol*. (2025) 3:72. doi: 10.1038/s44271-025-00251-4
- Hitchcock PF, Fried EI, Frank MJ. Computational psychiatry needs time and context. *Annu Rev Psychol*. (2022) 73:243–70. doi: 10.1146/annurev-psych-021621-124910
- Cioffi V, Mosca LL, Moretto E, Ragazzino O, Stanzione R, Bottone M, et al. Computational methods in psychotherapy: a scoping review. *Int J Environ Res Public Health*. (2022) 19:12358. doi: 10.3390/ijerph191912358
- Na H, Hua Y, Wang Z, Shen T, Yu B, Wang L, et al. A survey of large language models in psychotherapy: current landscape and future directions. In: *Findings of the Association for Computational Linguistics: ACL 2025*. Association for Computational Linguistics, Vienna, Austria (2025). p. 7362–76. doi: 10.18653/v1/2025.findings-acl.385
- Weinhardt D, Plomecka MB, Tezcan IM, Eckstein M, Musslick S. Automated discovery of sparse and interpretable cognitive equations. *PsyArXiv*. (2025). doi: 10.31234/osf.io/v86q5\_v1
- Musslick S, Bartlett LK, Chandramouli SH, Dubova M, Gobet F, Griffiths TL, et al. Automating the practice of science: opportunities, challenges, and implications. *Proc Natl Acad Sci U.S.A.* (2025) 122:e2401238121. doi: 10.1073/pnas.2401238121
- Abbass AA. *Reaching Through Resistance: Advanced Psychotherapy Techniques*. Halifax, NS: Seven Leaves Press (2015).
- Frederickson J. *Healing Through Relating: A Skill-Building Book for Therapists*. 2nd ed. Halifax, NS: Seven Leaves Press (2023).
- Frederickson J. *Clinical Thinking in Psychotherapy: What It Is, How It Works, and Why and How to Teach It*. New York, NY: Routledge (2025).
- McCullough L. *Treating Affect Phobia: A Manual for Short-Term Dynamic Psychotherapy*. New York, NY: Guilford Press (2003).
- Niederlohmann E. A transdiagnostic conflict square algorithm: a four node computational framework for psychotherapy and functional diagnosis. *Front Psychiatry* (2026). 17:1687372. doi: 10.3389/fpsy.2026.1687372
- Molodynski A, Linden M, Juckel G, Yeeles K, Anderson C, Vazquez-Montes M, et al. The reliability, validity, and applicability of an English language version of the Mini-ICF-APP. *Soc Psychiatry Psychiatr Epidemiol*. (2013) 48:1347–54. doi: 10.1007/s00127-012-0618-6
- Lieder F, Shenhav A, Musslick S, Griffiths TL. Rational metareasoning and the plasticity of cognitive control. *PLoS Comput Biol*. (2018) 14:e1006043. doi: 10.1371/journal.pcbi.1006043
- Lieder F, Griffiths TL. Resource-rational analysis: understanding human cognition as the optimal use of limited computational resources. *Behav Brain Sci*. (2019) 43:e1. doi: 10.1017/S0140525X1900061X
- Parr T, Pezzulo G, Friston KJ. *Active Inference: The Free Energy Principle in Mind, Brain, and Behavior*. Cambridge, MA: MIT Press (2022).
- Cheadle JE, Davidson-Turner KJ, Goosby BJ. Active inference and social actors: toward a neuro-bio-social theory of brains and bodies in their worlds. *Kolner Z Soziol Sozpsychol*. (2024) 76:317–50. doi: 10.1007/s11577-024-00936-4
- World Health Organization. *Ethics and governance of artificial intelligence for health: WHO guidance*. Geneva: World Health Organization (2021).